The Relationship Between Poverty and COVID-19

Tilman Bayer, Kieran Berton, Ruilan Liu, Toby Petty

4/13/2021

Introduction

The novel coronavirus which causes COVID-19 spreads easily and indiscriminately among humans in close quarters, requiring just a few respiratory droplets produced by an infected person to move on to its next host. However while the viral mechanism itself may act completely indiscriminately, there have been widespread concerns since the start of the pandemic that both the physical effects of the disease and the attendant economic devastation have disproportionately impacted the lower socioeconomic segments of society. In June 2020, UN chief António Guterres warned the General Assembly that the impacts of the COVID-19 pandemic were falling "disproportionately on the most vulnerable: people living in poverty, the working poor, women and children, persons with disabilities, and other marginalized groups".[1]

There are many plausible mechanisms that could be contributing to socioeconomic disparity in COVID-19 outcomes. Early in the pandemic there were stories of wealthy neighborhoods in metropolises like New York City emptying out as the better-off fled dense urban areas to escape the disease[2], whereas those without the financial security to be able to isolate, or the type of jobs conducive to working from home (the 'working poor'), had to continue working low-wage jobs throughout much of the lockdown period, thus increasing their chances of contracting the virus[3] [4]. Socioeconomic status and poverty levels are also closely intertwined with various other demographic characteristics of the population, such as race and ethnicity, education level, occupation types, and housing status; all of which have been linked to disparities in the impacts of COVID-19[5]. A Time article in March 2020–early on in the pandemic–identified a bewildering number of different ways in which the poor were uniquely predisposed to suffer its effects, including: being more likely to be uninsured or in prison, less able to stock up supplies to self-isolate, less likely to have access to a General Practitioner (GP), more likely to work in jobs that don't offer sick leave, more reliant on school-provided meals, less able to afford child-care, more likely to live in small, crowded spaces, etc.[6]

Numerous academic studies have already been conducted and found results suggesting that COVID-19 disproportionately negatively impacted the poor; for example finding much higher rates of infection in the poorest 100 counties in the US relative to the the richest 100 counties[7], or finding that those with COVID-19 symptoms were more likely to have lower incomes and less education[8], or finding that measures taken to promote social distancing were more effective at curbing infections in high income counties than in low income counties[9]. For our research question, we address the issue of poverty at the state level, examining whether there are significant differences in states' COVID-19 outcomes related to their levels of poverty. There are large differences overall in states' poverty levels, ranging from 7% in New Hampshire to almost 20% in Mississippi[10]. We aim to explore if negative COVID-19 outcomes, specifically death rate, is linked to

[1] https://news.un.org/en/story/2020/06/1067502
[2] https://www.nytimes.com/interactive/2020/05/15/upshot/who-left-new-york-coronavirus.html
[3] https://www.bbc.com/news/health-56334982
[4] https://www.cnbc.com/2020/09/19/coronavirus-how-the-pandemic-impacts-americas-working-poor.html
[5] https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html
[6] https://time.com/5800930/how-coronavirus-will-hurt-the-poor/
[7] https://www.frontiersin.org/articles/10.3389/fsoc.2020.00047/full#h4
[8] https://medicalxpress.com/news/2020-09-americans-sick-covid-disproportionately-poor.html
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7756168/
[10] https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_poverty_rate

the overall poverty level in a state.

Hence, our research question explores the relationship between pre-COVID measures of the socioeconomic makeup of each state and its COVID-related death rate as measured in 2021. We explore the relationship between the two metrics, seeking to understand the extent to which levels of poverty contributed to a state's COVID-19 death rate. We do not attempt to establish a formal causal relationship, because notions of causality around poverty are too complex and politically fraught (and because poverty levels are closely related to other variables we will consider, often with disputed causal directions), but we make mention of several possible causal pathways related to poverty and its effects. In addition to the main variable of the percentage of residents that are considered under the poverty line in each state, we also consider additional variables that we reason are important in trying to explain the variance in COVID-19 death rates between states, specifically: age, education levels, population density, housing density, and internet connectivity.

Data Description

To operationalize our research question, we utilized data from two main data sources, the New York Times' *Coronavirus (COVID-19) Data in the United States* dataset and the United States Census' *2019 American Community Survey (ACS).*

The New York Times has regularly released a series of data files with cumulative counts of coronavirus cases in the United States at the state and county level since late January 2020. The data was made publicly available for broad, noncommercial public use including by medical and public health researchers, policymakers, analysts and local news media. The data is compiled from state and local governments and health departments in an attempt to provide a complete record of the ongoing outbreak, and begins with the first reported coronavirus case in Washington state on Jan. 21, 2020. Dozens of NYT journalists work across several time zones to monitor news conferences, analyze data releases, and seek clarification from public officials on how they categorize cases to create this real time data. When the information is available, patients are counted where they are being treated, not necessarily where they live. For example, when a resident of Florida died in Los Angeles, her death was recorded as having occurred in California rather than Florida, though officials in Florida counted her case in their own records. As a result, the NYT data doesn't exactly match information reported by states and counties in some cases.

For our research purpose, we use the *state-level dataset* [11], in which case numbers and death numbers are provided in a time-series manner for each state, where deaths represents individuals who have died and meet the definition for a confirmed COVID-19 case; non-COVID-19 deaths like homicide, suicide, car crash or drug overdose were removed among confirmed cases with unambiguous information.

Our second primary data source, the ACS[12], is a nationwide survey that collects and produces information on social, economic, housing, and demographic characteristics about the United States population every year. This information provides an important tool for communities to use to see how they are changing, and for the purposes of our analysis the dataset provided important information on state-level poverty, education, and internet access levels, as well as other relevant community demographic information. The dataset also included measures from Puerto Rico and the District of Columbia.

The Census Bureau selects a random sample of approximately 3.5 million addresses to be included in the ACS, and each address receives a questionnaire via mail. This is a small number of households considering there are more than 140 million eligible addresses in the United States, and an address that receives ACS instructions will not likely find a neighbor or friend who has also received them. The sample is designed to ensure good geographic coverage and does not target individuals. Approximately 250,000 surveys are sent each month of the year, and results are then aggregated to create period estimates, which represent the characteristics of the population and housing over a specific data collection period. For our purposes, the

[11] The New York Times. (2021). Coronavirus (COVID-19) Data in the United States. Via https://github.com/nytimes/COVID-19-data, downloaded on March 27, 2021 from https://raw.githubusercontent.com/nytimes/covid-19-data/eab41bf87e3fed34ff31f30b2a1cc1ec73377eff/us-states.csv

[12] https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf

2019 ACS release was the most relevant year of data, as we wanted to isolate the demographic and community characteristics of each state before the pandemic from the effects of COVID-19.

Exploratory Data Analysis

Death rate To make a fairer comparison between states, we first converted the cumulative death number into death rate, or deaths per 100,000 people (`deathsperlakh`), by dividing by the population of each state, which came from the ACS dataset. We used cumulative counts of deaths and infections from the available data (on 2021-03-26) and restricted our analysis to only this static data. After removing Puerto Rico and other US entities (see below), there were 57 variables and 50 observations left in the NYT data. The mean death rate per 100,000 for all 50 states is 153.9, with standard deviation ± 59.7; confirming a lot of variance in the death rate between different states, which increases our confidence in its suitability as the dependent variable in our model.

Poverty rate To define whether a family or household lied below the poverty line, we relied on the US Census' definitions, which is adjusted yearly and depends on the size of a household, the number of children (if any), and the age of the head of the household. Estimates for the percentage of the total population of state residents below the poverty line (`percent_in_poverty`) ranged between 7-20% across the 50 US States and Washington D.C., but an outlier data point of 43% was reported for Puerto Rico. As this number was far outside the range of other states' reported values (see plot below) and our primary interest was analyzing the effect U.S. States' sociodemographic and socioeconomic characteristics had on their residents' rates of death to COVID, we chose to exclude Puerto Rico from our analysis. We also excluded all other US territories from our analysis, including Guam, the Northern Mariana Islands, and the U.S. Virgin Islands as well as Washington D.C. which is considered a federal district, not a state. Following these exclusions, the mean poverty percent for all 50 states is 12.1, with standard deviation ± 2.7.

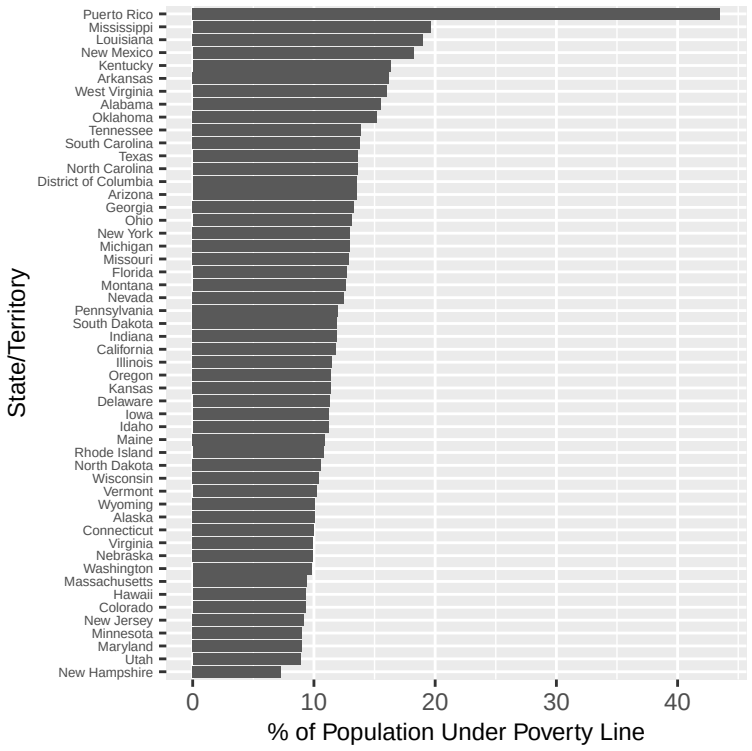

Figure 1: Percent of Population in Poverty by State/Territory

We initially examine just the relationship between poverty and death rates as defined above, to assess the validity of our research question; the scatter plot below (Figure 2) shows this relationship. On inspection it displays several characteristics that show promise. First, there appears to be a slight linear relationship from bottom-left to top-right, i.e. that higher poverty levels are somewhat correlated with a higher death rate. Secondly, the two states identified previously as having the lowest and highest poverty rates (Mississippi and New Hampshire) also appear at the low/high ends of death rate respectively, which further confirms the relationship. Thirdly, a group of states that does not appear to be closely following the linear relationship in the top-left quadrant of the plot are all regionally clustered in the North-East, suggesting that other variables on which these states share similar measures may be able to explain some of the deviations from the linear relationship (we discuss later that regional clustering of states may also present issues for the assumption of independence).

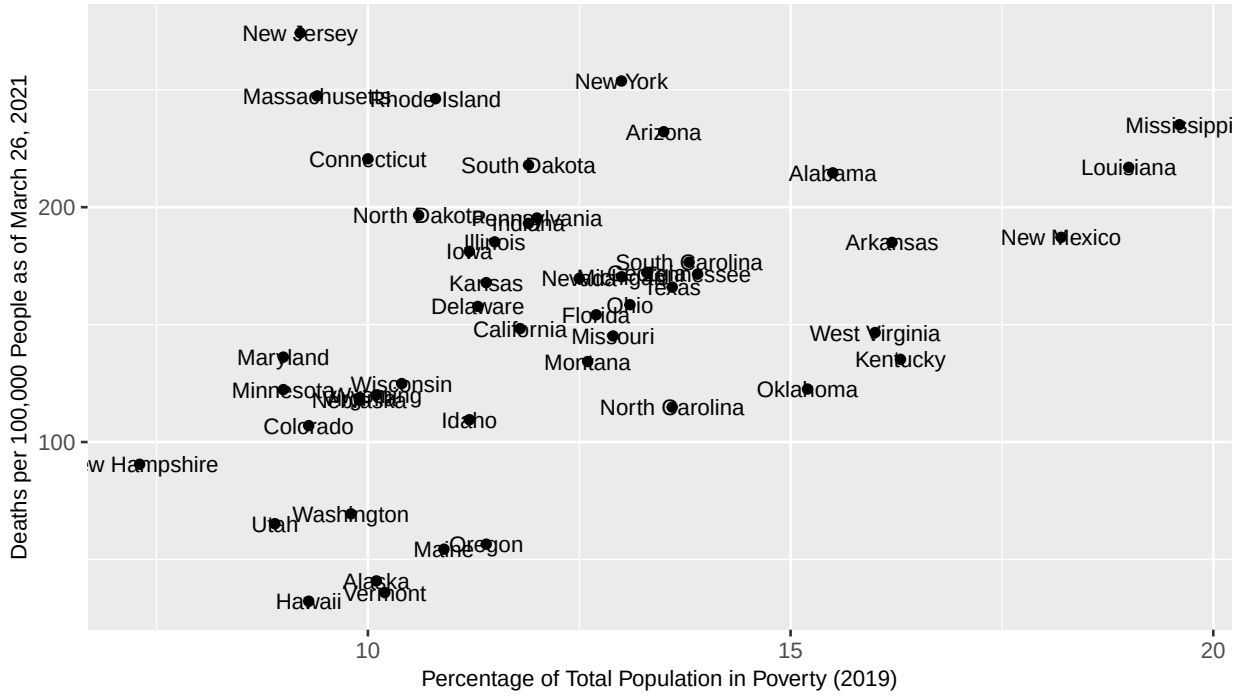

Figure 2: State Poverty and Covid Death Rate

Connectivity We considered two potential features relating to connectivity (percent without access to a computer as `wo_computer_pct`, and percent without access to the internet as `wo_internet_pct`), believing that it may have some effect on how COVID-19 impacts people, reasoning that people who are more connected would be more likely to have up-to-date information on the spread of the disease, prevention measures, testing and treatment information, etc. On inspecting the variables however we find that both are strongly correlated with overall levels of poverty (≈ 0.74 correlation with `wo_computer_pct`, blue scatter plot, and ≈ 0.85 correlation with `wo_internet_pct`, red scatter plot), and the model selection process confirmed that it was not adding enough novel information to improve model performance, so it was removed from Model 2. Furthermore, as the two variables are so highly correlated with each other (≈ 0.9 correlation; purple scatter plot below), we opted to include only `wo_internet_pct`, reasoning that in a pandemic having internet access is more important than computer access since people can also use phones or other devices to get information, but might not use a computer to access the internet.

Educatedness As part of our analysis, we also considered a number of measures of a state's level of education, reasoning that levels of education may often determine whether someone works as a 'white-collar' professional who is more likely to have been able to escape the spread of the disease through transitioning to

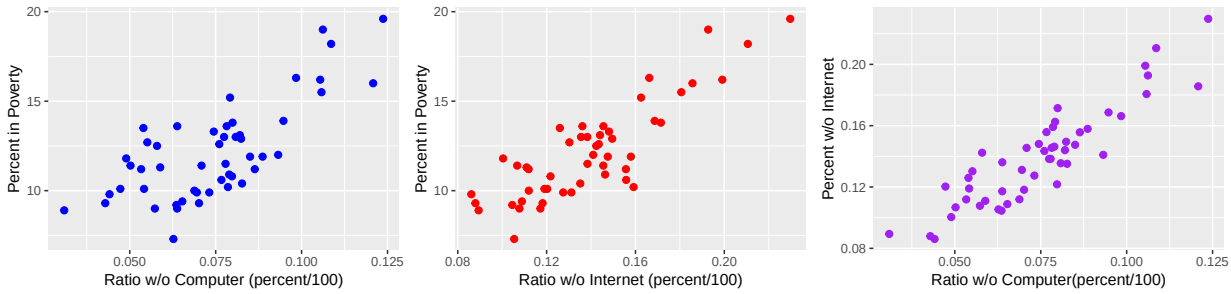

Figure 3: Three measures of connectivity, vs. poverty

working from home, as opposed to a frontline worker whose job cannot be done remotely, e.g in a 'blue-collar' manual labor job. We also speculate there may be additional reasons why being more highly educated may indirectly lead to different outcomes, such as being more discerning/engaged with COVID-19 news information and less likely to believe pseudo-scientific claims or 'fake news' stories shared through word of mouth or online, and therefore more willing to engage in behaviors that promote community health such as wearing a mask or social distancing. Estimates from the ACS dataset were used to operationalize the percentage of state residents above 18 and separately state residents above 25 that had reached various levels of education. The levels of education we considered were 1) completion of a high school degree (GED or equivalent), 2) some collegiate or post-secondary education, 3) completion of an associate's degree, 4) completion of a bachelor's degree, and 5) completion of a graduate degree or similar. For each of these of these variables we examined its correlation with the percentage of state residents in poverty. Plotted below are the two variables with the largest correlation with this metric (from left to right, Percent of Population (25+) with at Minimum a High School Degree and Percent of Population (18+) with at Minimum Some College Experience with correlations of ≈ -0.32 and ≈ -0.31 respectively). Other variables were removed from consideration for Model 2 due to both their similarity with the two metrics below and their less pronounced correlation with our primary covariate.

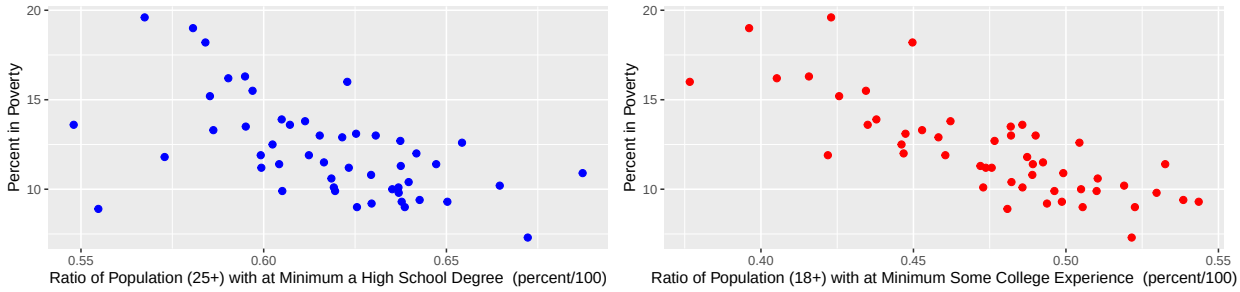

Figure 4: Two measures of educatedness, vs. poverty

Age It is well-documented that COVID-19 outcomes are markedly different for different age groups. For example, the CDC advises that those aged 85 and older are 8,700 times more likely to die from COVID-19 than those aged 5-17 years old, and still 2.7 times more likely than 75 to 84-year-olds (as of February 2021).[13] Therefore since our outcome variable is deaths we felt it important to include a variable which attempts to capture some of the variance in COVID-19 deaths attributable to age. The ACS data source breaks down state populations into a number of age-ranges, generally in buckets of 10 years, but since COVID-19 deaths are particularly prevalent in older age groups, we focused on those and examined the categories for those aged 65-74, 75-84, and 85+. Prioritizing model interpretability we decided to include only a single variable representing age, based also on our assessment that older age categories are strongly correlated; e.g. states

[13] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html

with a larger percentage of the population aged 85 or older tend to also have a larger percentage of the population aged 75-84. We calculated sums of these source ACS variables to create 3 variables representing percent of the population aged 65+, 75+, or 85+. All three variables were tested in the exhaustive search of all possible models, and having found that the 85+ category (`pct_age_85_plus`) produced the best results in terms of our chosen metric of Adjusted R^2 (unsurprisingly, given the CDC death statistics cited above), we include this variable as our age indicator.

Other Variables We additionally included several measures of population density, one being a measure of the number of state residents per square mile (`pop_den`), the other being a measure of the total number of housing units for each state resident (`house_per_capita`). Our reasoning for including these variables was based on the guess that more densely populated states would be more likely to have higher coronavirus infection rates due to the sheer number of citizens that could come into contact with one another. Similarly, larger households we hypothesized would also pose a higher infection risk. The housing data came from the ACS and estimates for total land area of each state came from the US Census Bureau's table of State Area Measurements and Internal Point Coordinates [14].

Anomalies and transformations We examined all of these variables for anomalies, finding none (except for the outlier of `percent_in_poverty` for Puerto Rico, which, as discussed above, is likely valid data but nevertheless contributed to our decision to remove it from the dataset).

We also visually inspected the histograms of all these variables to see if any of them might benefit from transformations (see the diagonal in Figure 5, "Feature Characteristics"). The only variable where the histogram indicated that a transformation should be considered was population density (`pop_den`), where the distribution was skewed towards zero, motivating us to try a log transformation (`log_pop_den`). However, the transformed variable ended up performing worse in the model selection phase (see below).

Model Selection and the Subset Selection Algorithm

After completing our EDA, the variables selected to operationalize the socioeconomic and sociodemographic characteristics of interest of each state were as follows:

Covariate	Variable Name
Percent of State Residents in Poverty	percent_in_poverty
Percent of State Residents without an Internet Subscription	wo_internet_pct
Density of State Population (in units of people/mile^2)	pop_den
Density of State Population (in units of people/mile^2), log transformed	log_pop_den
Total Housing Units per State Resident	house_per_capita
Percent of State Residents 25 Years or Older with at Minimum a High School Degree	hs_or_higher_25_plus_percent
Percent of State Residents 18 Years or Older with at Minimum Some College Experience	higher_edu_18_plus_per_capita
Percent of State Residents 85 Years or Older	pct_age_85_plus

The plot below (Figure 5) shows the crosswise correlation values between each variable as well as scatter plots showing the distribution of state-level data for each pair of variables (with the exception of `log_pop_den` which we ended up not using for other reasons, see below).

In order to select the best combination of these variables to include in our model, we used a Subset Selection

[14] https://www.census.gov/geographies/reference-files/2010/geo/state-area.html

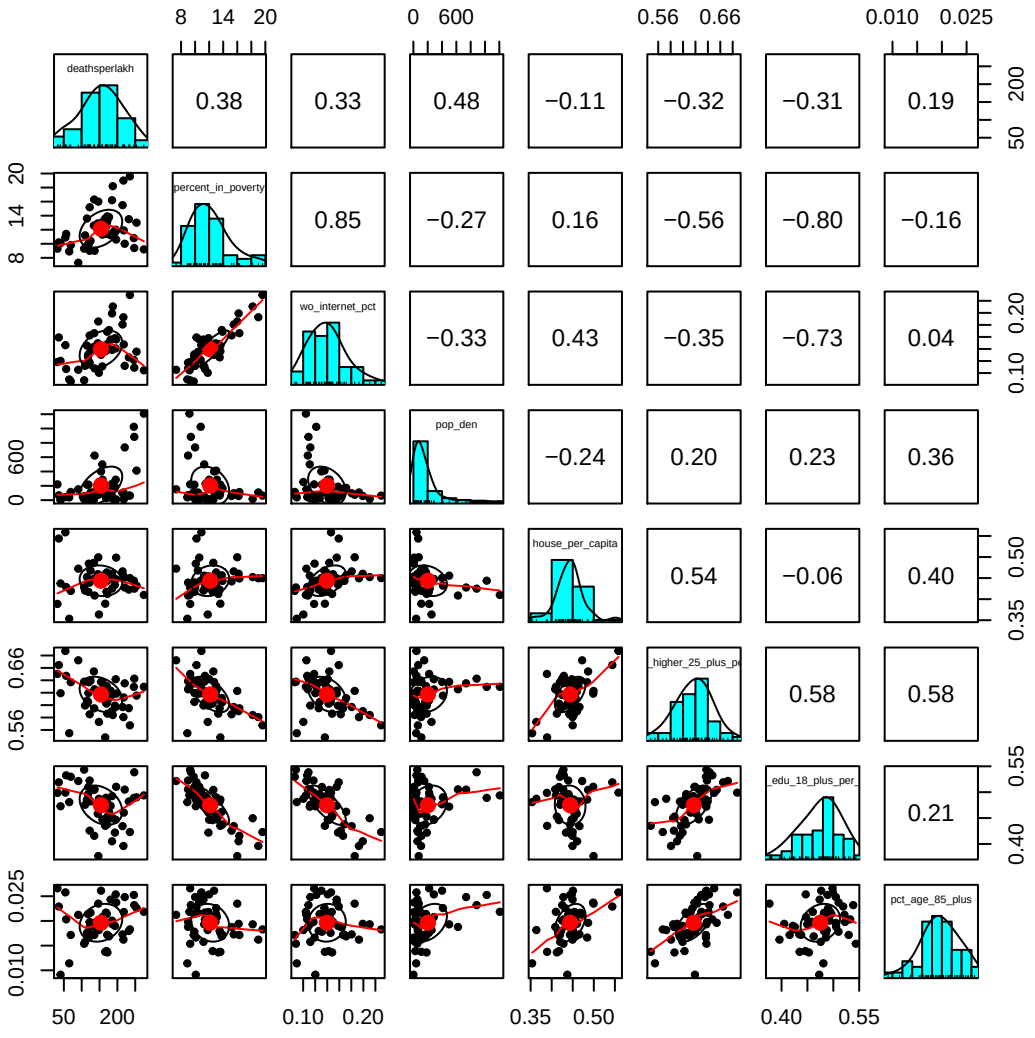

Figure 5: Feature Characteristics

Algorithm[15] to identify a subset of the predictors that we believe to be the best at explaining the response. We chose to use Adjusted R^2 as the metric to choose between models, since it is easily interpretable as it does not depend on the scale of the variable (as for example MSE does); the metric allows for direct comparison between models with different features as it accounts for the number of features in the models (whereas R^2 does not). The Subset Selection algorithm works by comparing R^2 values for every possible combination of input variables, in other words fitting every unique combination of 1 variable, then 2 variables, and so on to a specified maximum number of input variables, and reporting which combination of input variables for each number of possible inputs reported the highest R^2 value. The second stage of the algorithm then compares the Adjusted R^2 values for the 'best' model at each number of possible inputs. We specified that the algorithm should always retain the coviarate representing the percent of state residents in poverty as this is our primary variable of interest.

We also ran the algorithm with the `log_pop_den` variable included instead of its non-transformed original, which performed notably worse (adjusted R^2 of ≈ 0.36). When we included both the transformed and non-transformed metrics, the 'best' model yielded by the algorithm did not use `log_pop_den`. Because of this and additional concerns that a model using both as input variables would be hard to interpret, we decided to exclude `log_pop_den` from the further model building process.

Generate Models

Using this method, the algorithm identified that the model with the greatest explanatory power as judged by the R^2 and Adjusted R^2 values included only five covariates:

Covariate
Percent of State Residents in Poverty
Density of State Population (in units of people/mile^2)
Total Housing Units per State Resident
Percent of State Residents 25 Years or Older with at Minimum a High School Degree
Percent of State Residents 85 Years or Older

This algorithm reported an R^2 of 0.604 and an Adjusted R^2 of 0.559. In the table below, we include statistical outputs for three models. Model 1 is a limited model with only our primary covariate included, Model 2 is the model chosen by the subset selection algorithm and is the focus of our analysis, and in Model 3 we included all other covariates that were removed by the algorithm.

[15] See Section 6.1 in *An Introduction to Statistical Learning* by James, Whitten, Hastie, and Tibshirani https://www.statlearning.com/

```
##
## Model Comparison
## ================================================================================================
##                                                Dependent variable:
##                                    ------------------------------------------------------------
##                                                     deathsperlakh
##                                          (1)                (2)                    (3)
## ------------------------------------------------------------------------------------------------
## percent_in_poverty                     8.383***           4.986                  3.750
##                                         (2.958)           (3.387)                (5.043)
##
## wo_internet_pct                                                                   402.277
##                                                                                  (494.890)
##
## pop_den                                                    0.140***               0.144***
##                                                            (0.026)                (0.027)
##
## house_per_capita                                           370.263                256.142
##                                                           (264.758)              (324.030)
##
## hs_or_higher_25_plus_percent                            -1,242.771***          -1,190.913**
##                                                           (417.487)              (469.612)
##
## higher_edu_18_plus_per_capita                                                     198.499
##                                                                                  (268.983)
##
## pct_age_85_plus                                          4,023.404*             3,458.614
##                                                         (2,024.794)            (2,171.352)
##
## Constant                               52.114             588.706***             482.254**
##                                       (36.775)           (207.139)              (238.106)
##
## ------------------------------------------------------------------------------------------------
## Observations                             50                  50                     50
## R2                                      0.143               0.604                  0.613
## Adjusted R2                             0.126               0.559                  0.549
## Residual Std. Error              55.864 (df = 48)    39.675 (df = 44)       40.136 (df = 42)
## F Statistic                   8.035*** (df = 1; 48) 13.419*** (df = 5; 44) 9.508*** (df = 7; 42)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

The three coefficients in model 2 that were statistically significant (`pop_den`, `hs_or_higher_25_plus_percent`,`pct_age_85_plu
we discuss further in the conclusion.

CLM Assumptions and Limitations of the Model

Below we discuss several of the assumptions of the Classical Linear Model (CLM) and whether our model satisfies them.

IID Given the broad nature of our observations (i.e. 1 per state) we acknowledge that there may be several issues with the assumption that our data was drawn independently from an identically distributed population (IID), which may diminish the ability of our final model to fully answer the research question as originally posed, or may result in greater uncertainty than the model's standard errors suggest. In considering the IID assumption it is first useful to explicitly identify the population from which our sample is drawn, and about

which it is supposed to make inferences. The population is effectively all people in the United States, and the 'sampling' in our research is just the aggregation of the population into the somewhat arbitrary geographical and political construct of 'states'. Viewed in this way it is fairly obvious that this poses some clear challenges to the assumption of *independence*:

- States are often grouped into wider regions (such as 'New England', the 'Midwest', 'the South', etc.), which are supposed to represent some set of shared geographical, political, social, cultural, economic, or demographic features that broadly characterizes that region's inhabitants as distinct from the rest of the states[16]. If this is the case then we do not really have independence of states, since we may expect states drawn from the same region to also have similar characteristics in terms of the variables of interest in our research; i.e. poverty levels and COVID-19 impacts.
- States are also political entities, and since the government response to COVID-19 was largely marshaled at the state level, we may expect there to be some clustering of observations by political affiliation. For example, Wyoming and Alabama are geographically separate and different, but both are governed mostly by Republican politicians, so we may reasonably expect similarities both in how these states responded to the pandemic, and in how they generally perceive and respond to poverty.
- States are diverse in a number of different ways that may materially impact both the input and response variables:
  - It has been suggested that the spread of COVID-19 may be impacted by different weather or climatic conditions[17]; therefore we may expect clustering of states with similar climates, e.g. those that form the South-Western United States or the North-Eastern United States.
  - States vary greatly in both total populations and population densities, therefore there may be clustering of observations based purely on these physical characteristics of how people are distributed throughout a state; e.g. sparse rural states vs. small populous states. Our model attempts to control for some of these differences by including a variable for population density.

Another potential issue with the IID assumption for these observations is that the amount of time in which states have had COVID-19 has not been uniform: the first confirmed case in the United States was on January 20th 2020, but it was not confirmed in all states until mid-April[18]. The pandemic has been marked by dramatic surges and falls in cases in different regions of the US, and at the time of writing 32 states are characterized as "higher and staying high"[19] by the New York Times; therefore we cannot say that the response variable represents all COVID-19 related deaths a state will have (although we hope it at least captures the majority). Since most states have now had coronavirus cases for at least a year we believe this is enough time to capture meaningful differences in how states have been impacted by COVID-19. One way to deal with this issue more definitively would be to re-run the model with updated data once the pandemic has subsided.

These issues with the IID assumption are one reason why we need to be cautious with how we interpret the model's results, and do not overstate how they would generalize to the population. The clustering of states is problematic, but geographical and political clustering of observations would still be an issue with more granular data, e.g. at the county level. Ideally we would want a large random sample which was representative of the overall geographical distribution of the population, but without access to such we have to make do with the data we have available, and acknowledge that the standard errors the model produces will likely be lower than the true uncertainty measure.

Linear Conditional Expectation and Homoskedasticity A visual inspection of the plot of residuals vs. predicted values for Model 2 shows that the assumption of linear conditional expectation appears seriously violated. If it were satisfied, the blue line in the plot would be flat. But that is not the case; rather it indicates that the model tends to underpredict around a predicted value of 150 deaths per 100,000 (by roughly 25 on

[16] See e.g. https://en.wikipedia.org/wiki/American_Nations
[17] https://www.medicalnewstoday.com/articles/how-does-weather-affect-covid-19
[18] https://en.wikipedia.org/wiki/COVID-19_pandemic#North_America
[19] https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

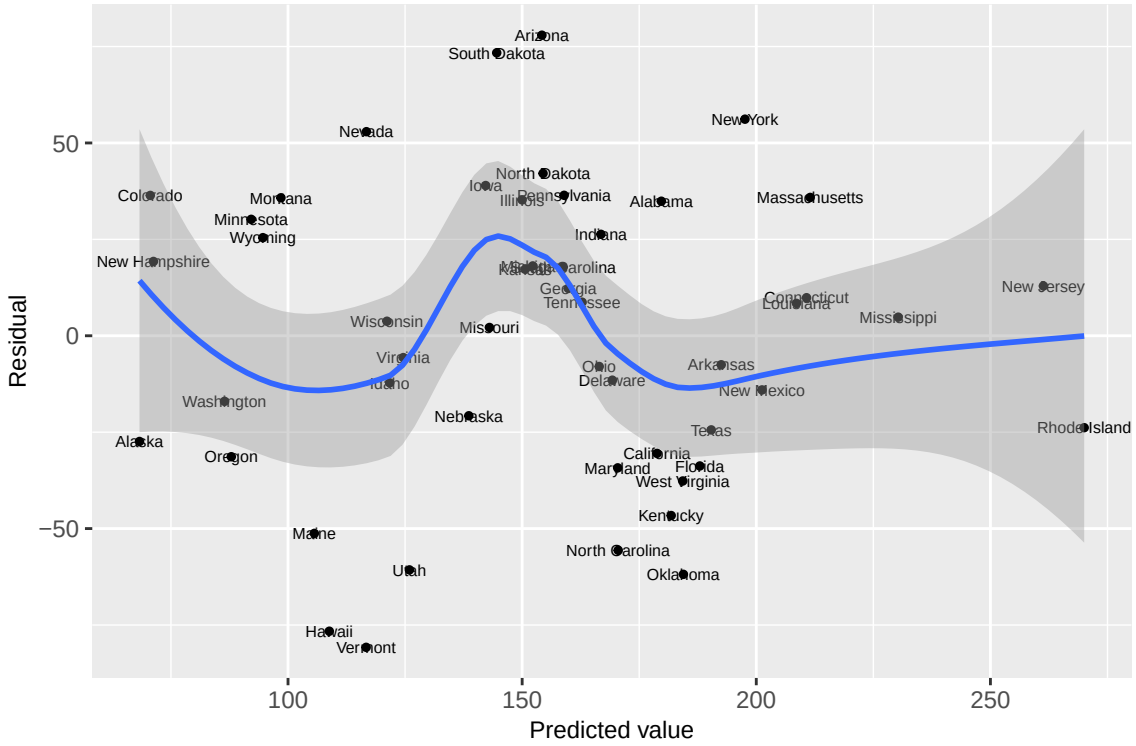

Figure 6: Residuals vs. Predicted values (Model 2)

average), whereas predicted values in the ranges of around 20-130 and above 170 are likely to be lower than the true value.

The same plot also allows us to examine the assumption of homoskedasticity. The width of the grey band in the plot does not stay the same over the range of predicted values, showing that this assumption is likewise violated, especially so on the right hand side (largest predicted death rates), where it reaches almost 100, compared to less than 40 near the center of the graph.

Normally Distributed Errors To assess whether the model's errors are normally distributed we examine a QQ Plot to see if there is a noticeable deviation from the theoretical normal distribution of residuals. On inspection we find no obvious evidence of deviations from normality that concerns us; there are some slight deviations from the perfect normal line at the upper and lower ends of the plot, but this is not unexpected given the small sample size.

Conclusion To conclude this section, we find violations already in the first group of the five CLM assumptions (specifically regarding IID and linear CE), meaning that we cannot assume that the estimators resulting from our model are unbiased. Furthermore, this also means that its estimates for the uncertainty in the model estimates cannot be assumed to be unbiased.

Omitted Variables Discussion

The main relationship we explored in our model was between a state's overall level of poverty and its COVID-19 death rate. Poverty is an extremely complicated topic with complex relationships with many of the variables we considered, in which directions of causality are often unclear or even politically disputed. For example, there is a correlation of ≈ -0.56 between a state's percentage in poverty and the percentage of its population (25+) with a high school degree (at minimum). In terms of causality we could easily argue

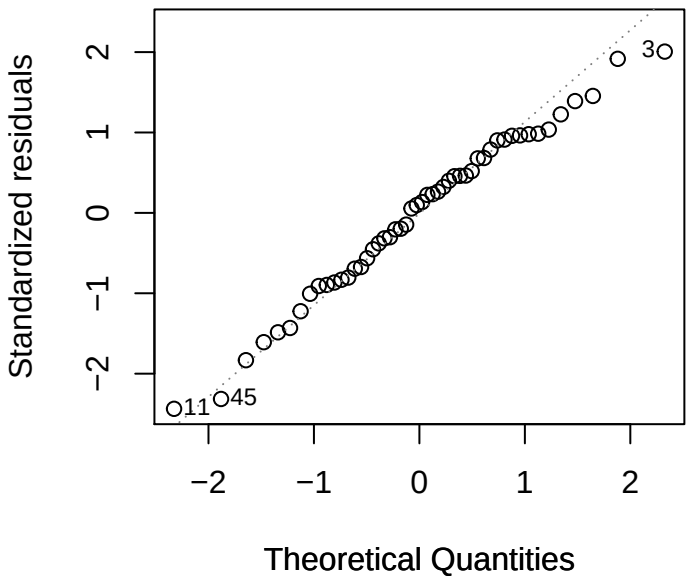

Figure 7: Normally Distributed Errors

that being poor might lead to one being more likely to be uneducated, and likewise we could argue that the benefits of education might lead to one being more likely to be out of poverty. Similar bi-directional causal arguments could be made for the relationship between poverty and several of the variables we considered (e.g. age, access to the internet, etc.) We do not suggest that a simple linear model could capture such a tangled web of causality, and hence we do not attempt to determine specific omitted variables that may causally introduce large biases to our model. However, in terms of additional variables which we didn't include that we believe may improve our model's performance in explaining the variance in COVID-19 deaths between different states, we do suggest two particular factors that we would seek to explore further in any follow-up to this study:

- *Politics* - COVID-19 became a highly contentious political issue in the United States, particularly as the pandemic occurred in the midst of a hugely divisive Presidential election (and may indeed have caused some of that divisiveness). Broadly speaking the popular perception of political responses to the pandemic is that Republican-led states were more opposed to restrictions (e.g. lockdowns) and mitigating behaviors (e.g. mask-wearing), whereas Democratic leaders were more embracing of such measures. We would seek to introduce a variable that could characterize overall political leanings of the state, e.g. the political party of its governor.
- *Weather* - As with seasonal flu, COVID-19 outcomes have been linked to climatic conditions[20], so we would seek to introduce a variable that captured some of this variation, such as annual average temperature.

[20] see *IID* Section above

Conclusion

As indicated in the model results, Model 2 shows the highest Adjusted R^2 of 0.56 and the lowest Residual Standard Error (RSE) of 39.68 among all three models. By adding two more variables, the Adjusted R^2 decreased by 0.01 and the RSE actually increases, which further justifies our selection of model. For context, according to one statistician, "any study that attempts to predict human behavior will tend to have R-squared values less than 50%,"[21] which would mean that Model 2 performs quite well. We do however need to caution that this value, as all our conclusions in this section, depend on the assumption that the CLM violations that we found above do not impact the accuracy of our regression reasoning too negatively.

Unsurprisingly, `pct_age_85_plus` shows the biggest impact: If a state has an 1% increase in population of age 85 and plus, our model predicts that the COVID-related deaths would increase by 4023.4 per 100,000 people. This is consistent with the aforementioned existing research that identified aging as a critical factor contributing to COVID-related deaths, with 85+-year-olds having (as of February 2021) a COVID-19 death risk 8,700 times higher than that of 5-17-year-olds, according to the CDC.[22] Compared to the overall average death rate for all US states of 153.9 deaths/100,000 people, this appears to be of clear practical significance, as the mean percentage of residents age 85 years or older across all states was ≈ 2.

Education also shows a strong relationship with death rate: for states with 1% more people finishing high school education and above, the COVID-related deaths can be reduced by -1242.77 per 100,000 people. One of the possible explanations is that finishing higher education indicates that a person is more likely to do office work, which provides more flexibility to work remotely without exposing to the crowds. It also means that a person is more likely to be economically sustainable, or less likely to fall into poverty. This could partially explain why the `percent_in_poverty` impact is so small as it is absorbed by `hs_or_higher_25_plus_percent`, as discussed above. Compared to the baseline average of ≈ 62 of state residents 25 years or older having completed at minimum a high school degree, this result shows the coefficient for education is also practically significant.

Our primary research question was to explore the relationship between between poverty and death rate, and in Model 2 the `percent_in_poverty` variable actually shows insignificant impact on the dependent variable, and its coefficient is small relative to other coefficients for education, age, and housing per capita (although these measures are not all on exactly the same scale, so we are cautious not to infer too much from these coefficient sizes). We do not interpret this to mean that poverty levels have no relationship with COVID-19 outcomes. Firstly, in the baseline model the poverty variable does have an extremely significant ($p < 0.01$) relationship with COVID-19 deaths, but with an Adjusted R^2 of just 0.126. Secondly, as discussed previously, there may be complex causal relationships between poverty and many of the variables included in the second and third models (especially education). Both of these statements support the hypothesis the poverty played a large part in determining a state's COVID-19 outcomes, but that its influence is complicated and mediated through a variety of intermediary mechanisms. While we can observe poverty's (small) effect directly in Model 1 with statistical significance, the fact that in Model 2 education has a much larger coefficient and is more significant does not invalidate poverty as an explanatory variable, education simply does a better job of explaining the variance in COVID-19 deaths in the observations.

With the addition of two extra variables in Model 3 (`wo_internet_pct` and `higher_edu_18_plus_per_capita`) there are tell-tale signs that the model becomes overspecified and its results should not be trusted as much as those of Model 2. In Model 3 several of the standard errors in the variables from Model 2 inflate and become larger than the coefficients themselves (`percent_in_poverty`, `house_per_capita`), and this is also the case for the two new variables introduced. This means it becomes hard to determine whether the variables have any effect on the output at all. In Model 3 we have a good theory for understanding why these problems have occurred, which is that the two new variables contain information that is largely already in the model (see earlier discussion re their correlations with the `percent_in_poverty` and `hs_or_higher_25_plus_percent` variables), so we have effectively introduced redundant predictors which inflate the standard errors[23].

[21] https://statisticsbyjim.com/regression/how-high-r-squared/
[22] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html
[23] https://online.stat.psu.edu/stat501/lesson/10/10.1