MIDS W203 - Lab 2 - Research Proposal

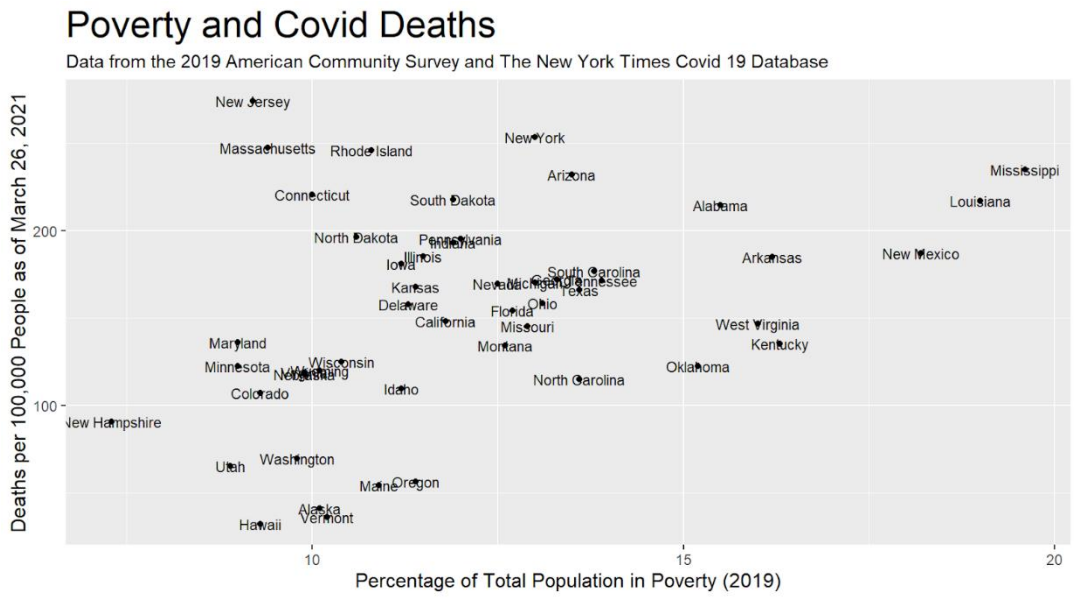
Authors: Tilman Bayer, Kieran Berton, Ruilan Liu, Toby Petty

Research Question

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”. 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 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.

Hence, we propose a research question that explores the relationship between pre-COVID measures of the socioeconomic makeup of each state and the COVID-related death rate in 2021. We are interested in exploring the causal relationship between the two metrics, but our focus will be primarily on isolating the explanatory effect of levels of poverty on COVID death rate. We anticipate there will be difficulty establishing a formal causal relationship in our model, but will attempt to identify any causal pathways that help to explain the output of our model. As a key variable for our limited model, we envisage using the percentage of residents that are considered under the poverty line (or a similar poverty metric) in each state.

Our investigation of available datasets and initial exploratory data analysis suggest that this research question is a feasible one. See e.g. the preliminary scatterplot of the two aforementioned variables, below which shows some limited relationship between states with high poverty levels and states with high COVID death rates. We will primarily use data from the American Community Survey and the New York Times’ COVID-19 dataset. Apart from the poverty percentage used in the limited model, we anticipate that other key variables will include age (operationalized as e.g. the percentage of the population above a threshold like 75 years that indicates strongly elevated COVID-19 fatality risk), and education attainment (operationalized e.g. as the percentage of the population holding a high school degree). As we progress, we will also consider the relationship between our key variables and population density, housing unit sizes, access to healthcare facilities or health insurance, and other potential factors.



Plan of Actions

We will split into two groups working on below items and write up the findings together:

1. Describe/summarize the available datasets
2. More closely example the distributions of the planned variables (e.g. using Q-Q plots or histogram), in particular to check for anomalies and to validate linear regression assumptions
3. Finalize selection of appropriate variables, both for the core model (model 2) and the extended one (model 3)
4. Construct the three regression models
5. Explain regression results
6. Evaluate whether adding variables in model three improve our understanding of the research question
7. Discuss possible problems of the models (variables collinearity, omitted variables)
8. Finalize report

Data Sources

- American Community Survey https://data.census.gov/cedsci/table?q=ACS&g=0100000US.04000.001&tid=ACSDP1Y2019.DP05&moe=false&hidePreview=true
- State Area Measurements and Internal Point Coordinates https://www.census.gov/geographies/reference-files/2010/geo/state-area.html

New York Times COVID-19 data https://github.com/nytimes/covid-19-data