Analyzing the Relationship of Economic Indicators on COVID-19 Outcomes

Michael Lewis[*] Your Name2[†]

2022-11-28

Abstract

This is our informative abstract of fewer than 200 words. It describes what we investigate, how we investigate it, and what we find.

1 Introduction

In this section, we introduce the reader to the phenomenon we investigate. We describe the way in which our analysis contributes to an important intellectual debate, or how it answers a pressing political or social question. We introduce our hypotheses, data, and results. We signpost for the reader what's coming in the rest of the paper.

We remember that our paper is not a mystery novel. We note our core results early and often.

Throughout our paper, we use active, first-person language and avoid the passive voice. For example, we write "we examine the relationship between X and Y"; we do not write "the relationship between X and Y was examined." Where we do the analysis, we speak about it transparently.

2 [Our Substance and Context Section Title Here]

Here we go deeper into the intellectual debate, the political and social context of our investigation. To give the reader a clear sense of why we are writing this paper, we describe the relevant scholarly, technical, or popular literature. We give this section a meaningful *substantive* title; it is not entitled "Literature Review", for example. We cite at least three published, peer-reviewed scholarly works. For example, we could cite Moore and Reeves (2020) or Moore and Ravishankar (2012), which we discussed in class.[1] We only cite others' work in our paper when it enhances the reader's understanding of what we, the authors of this paper, are doing. We connect everything we cite to *our* investigation; this is our original research, not a book report or an annotated bibliography.

In order to integrate citations into the References section below, we add entries into our file `main.bib`. This is a plain-text file that we edit in RStudio. We store `main.bib` in the same folder as our paper's `.Rmd` and `.pdf` files. Its entries are formatted so that they can be knit to `.pdf`; see https://j.mp/2UzTXEZ for example entries for articles, books, and miscellaneous. We can get these entries automatically from Google Scholar by turning on BibTeX in the Google Scholar Settings - Bibliography Manager. Perhaps we use a tool like free, open-source BibDesk to help us manage the `.bib` file.

[*] SIS/CAS, American University

[†] American University

[1] To cite a paper within parentheses, use, e.g., (Moore 2012).

3 Data and Methods

3.1 Data Sources

This project sources two datasets in order to investigate the relationship between certain economic and social variables American Community Survey (ACS) and COVID-19 data in the DC, Maryland, and Virginia. The American Community Survey (ACS) is acquired from https://api.census.gov/data/key_signup.html and COVID-19 Data is acquired from https://apidocs.covidactnow.org.

Our final combined data set, with both ACS and COVID-19 data, has 158 total observations with 13 variables of interest.

The American Community Survey is a perpetual survey run by the United States Census Bureau that provides the United States government, and its population, access to information surrounding different economic and social characteristics such as jobs, income, and education on a yearly basis. The ACS contacts over 3.5 million households every year for participation in this survey, where all individuals contacted are legally obligated to answer all the questions in the survey. The sample is selected at random by the Census Bureau, where no address is chosen more than once every 5 years.

Data surrounding the spread and impact of COVID-19 in the United States was sourced from COVID ActNow. This organization is a collection of scientists, pandemic experts, engineers, and public health experts who collect COVID-19 data in the United States, and then subsequently assess its quality. Their works supports federal, state, and local government agencies and their data sets are considered some of the most-trusted regarding the COVID-19 pandemic.

In determining the variables of interest that we wanted to investigate from the American Community Survey, we were most interested in the number of working individuals aged 16 years and older that commute to work, the total mean household earnings, the number of individuals that have health insurance coverage, and the percent of working individuals aged 16 years and older that are employed in an essential worker capacity. The specific variables and their official ACS codes are below:

- DP03_0018 Estimate, **Commuting to Work**, Workers 16 years and over
- DP03_0065 Estimate, Income and Benefits (in 2021 inflation adjusted dollars), **Total households, With earnings, Mean earnings** (dollars)
- DP03_0095 Estimate, **Health Insurance Coverage**, Civilian noninstitutionalized population
- DP03_0042P Percent, Industry, Civilian **employed** population 16 years and over, **Educational services, and health care and social assistance** (Note: we classify this variable as "essential employees", however, we recognize that this term is imperfect in encapsulating varying definitions of "essential" across three states and 150 plus counties. Nonetheless, we believe that these occupations generally represent what can be considered "essential" occupations during the pandemic)

With the variables selected from the ACS, we first started out with time series COVID-19 data for Maryland, DC, and Virginia. We only selected the year 2021 for COVID-19 data in order to match the ACS data from 2021. We were able to get a daily cumulative case and death count from COVID ActNow, but in order to ensure that our analysis was on the same level as ACS data, we created two new variables – average cases from 2021 and average deaths from 2021. The specific variables of interest from this data set:

- Average Cases
- Average Deaths

For the purpose of building a model to assess the significance of the relationship between multiple economic and social indicators and the spread of COVID in DC, Maryland, and Virginia, we aimed to combine both data sets to prepare it for analysis. We were able to load in both Census ACS data and COVID-19 data from COVID ActNow through their respective APIs, and then subsequently cleaned the source data.

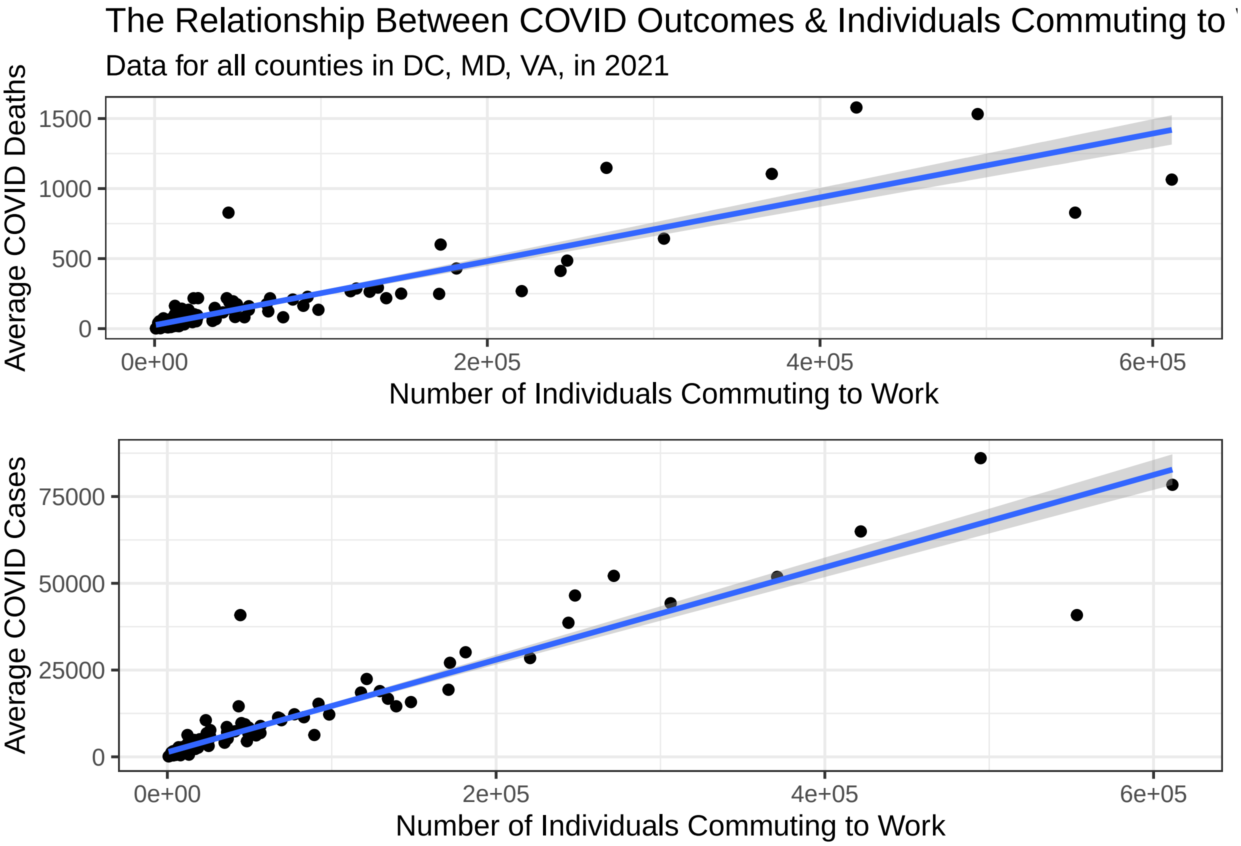
For cleaning the COVID-19 data, we parsed out DC, Maryland, Virginia from our complete time series data from COVID ActNow, then created a separate data set with just these three geographic locations, and computed average cases and deaths for each county during 2021. Then we merged both data sets to combine variables from both sets.

We conduct our analyzing using R version 4.2.2 (R Core Team 2022). To clean our data we primarily used the `tidyverse` (Wickham et al. 2019) packages as well as `janitor` and `lubridate`. Data collection was aided by the `tidycensus` package, which allowed us to pull direct data from the ACS/US Census Bureau. Finally, we used the the `keyring`, `httr`, and `jsonlite` packages for API imports from the respective sources.

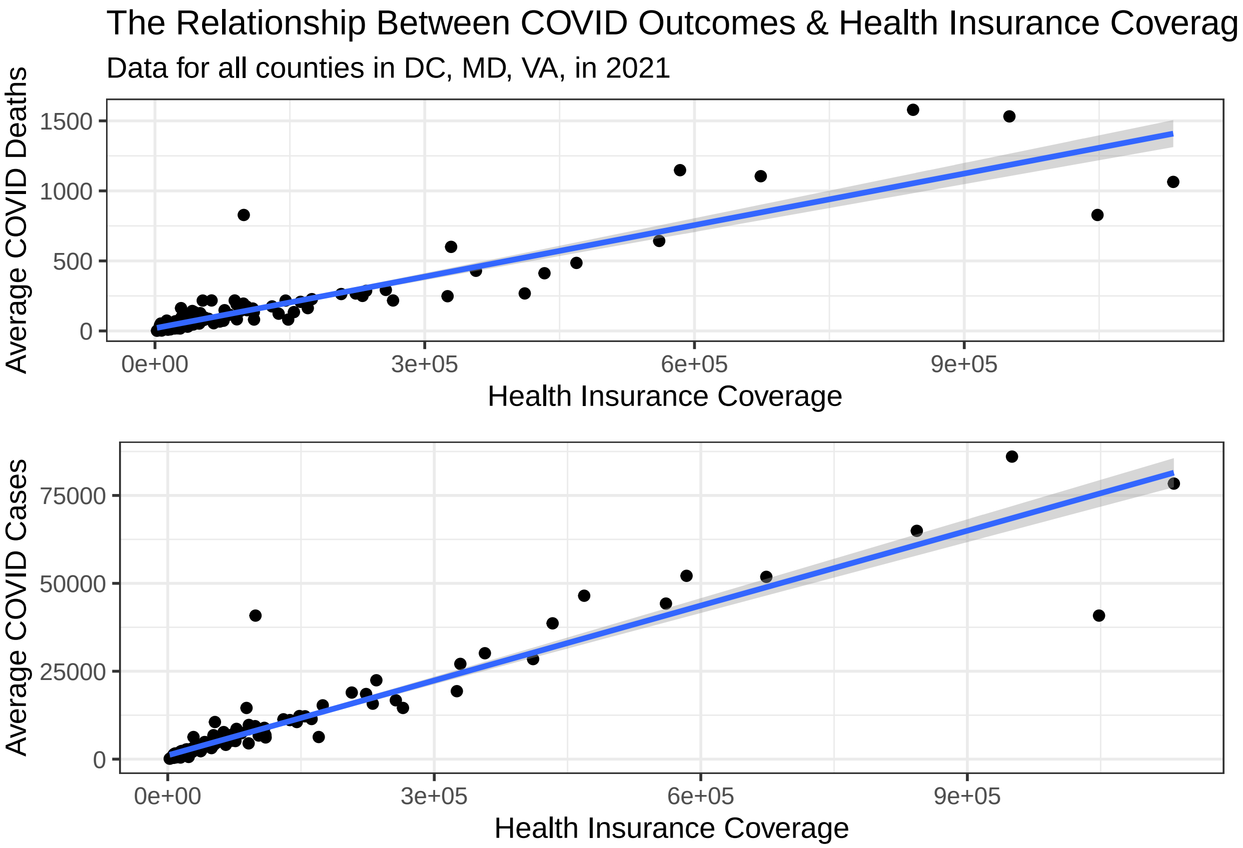
Since this project aims at investigating the relationship, and its subsequent significance, between certain economic and social indicators and COVID-19 average cases and deaths in different locales within DC, Maryland, and Virginia. Because of this goal, a linear model is the best fit to further analyze our research question.

3.2 Methods

We conduct a Ordinary Least Squares (OLS) regression model to explore the relationship between the aforementioned economic indicators and average COVID outcomes in 2021. We chose this method after examining the raw data and noticing patterns which indicated that a linear model may be appropriate. The below figure illustrates this phenomenon with individuals commuting to work and mean COVID deaths. While there is more variability in cases at the upper-ends, on balance the data follows a linear pattern.



Other variables also show evidence of a linear relationship. In our exploratory data analysis we found a positive association between health insurance coverage (or number of individuals in a given county with health insurance) and COVID deaths and cases. This finding seemed counter intuitive and was something we wanted to examine more closely after controlling for the effect of other predictors.



This exploratory data analysis led us to pursue linear regression as a method for analyzing multiple economic indicators and mean COVID outcomes. Specifically, we arrived at three research questions: (1) is health insurance coverage still positively associated with worse COVID outcomes even after controlling for the effects of other predictors, (2) what variables had the largest effect on mean COVID cases and deaths, and (3) how important is wealth in predicting COVID outcomes? The next section will discuss our model and results.

4 Findings

Here, we explain and interpret our results. We try to learn as much as we can about our question as possible, given the data and analysis. We present our results clearly. We interpret them for the reader with precision and circumspection. We avoid making claims that are not substantiated by our data.

Note that this section may be integrated into Section **??**, if joining the two improves the overall presentation.

Our results for the `cars` data include estimating the linear model

$$\text{Distance}_i = \beta_0 + \beta_1(\text{Speed}_i) + \epsilon_i.$$

Below we show the model estimates. The first table uses `xtable()`, the second uses `stargazer()` (Hlavac 2018).

Using the `cars` data, we find that each unit of speed is associated with 3.9 more units of distance. We draw out what this really means, and what it implies. For example, if a typical difference among our observations

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17.58	6.76	-2.60	0.01
speed	3.93	0.42	9.46	0.00

Table 1: Our Informative Caption

Table 2: Our Informative Title

	Outcome
	dist
speed	3.93***
	(0.42)
Constant	−17.58**
	(6.76)
Observations	50
R^2	0.65
Adjusted R^2	0.64
Residual Std. Error	15.38 (df = 48)
F Statistic	89.57*** (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

is 7 units of speed, then our model estimates that a typical difference in distance among our observations is $7 \times 3.9 = 27.3$ units of distance. We describe the substantive relevance of this number.

5 Discussion

We remind the reader what this paper was about, why it was important, and what we found. We reflect on limitations of the data or methods. If we have specific advice for someone picking up where we leave off, we provide that guidance. We avoid making trite statements like “more research should be done”.

References

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables.* Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Moore, Ryan T. 2012. “Multivariate Continuous Blocking to Improve Political Science Experiments.” *Political Analysis* 20 (4): 460–79. https://doi.org/10.1093/pan/mps025.

Moore, Ryan T., and Nirmala Ravishankar. 2012. “Who Loses in Direct Democracy?” *Social Science Research* 41 (3): 646–56. https://doi.org/10.1016/j.ssresearch.2011.10.003.

Moore, Ryan T., and Andrew Reeves. 2020. “Defining Racial and Ethnic Context with Geolocation Data.” *Political Science Research and Methods* 8 (4): 780–94. https://doi.org/https://doi.org/10.1017/psrm.2020.10.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing* (version 4.2.2). Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.