Analyzing the Relationship of Economic Indicators on COVID-19 Outcomes

Michael Lewis* Aarushi Sahejpal† Katherine Prince‡

2022-11-28

Abstract

Using an Ordinary Least Squares (OLS) regression model, this paper investigate the relationship between social, economic and health determinants and average COVID-19 cases and death using data from the American Community Survey from the US Census Bureau and COVID-19 timeseries data from COVID ActNow. We find that health insurance coverage is positively associated with COVID-19 cases and deaths, that counties with a higher percentage of essential workers are positively associated with an increase in cases and deaths. Finally, more individuals commuting and an increase in mean earnings are both associated with a decrease in cases and deaths. All of these predictors take into account the population at the county-level.

1 Introduction

This project investigates the spread and impact of COVID-19 among populations in DC, Maryland, and Virginia. But more specifically, it focuses on looking directly at how COVID-19 intersects with different social, economic, and health determinants that impact the American population. As to be noted in the subsequent sections, the COVID-19 pandemic's impact on the United States has, and is, well documented but there is a local lens that is missing from the analysis of its true impacts, hence the goal of this project.

Using data from the American Community Survey and COVID ActNow, through the mechanism of an Ordinary Least Squares (OLS) regression model, this project aims answering three research questions specifically for DC, Maryland, and Virginia: (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? We hypothesis that (1) health insurance coverage is negatively correlated with COVID-19 cases and deaths, (2) that counties with more essential workers and more individuals commuting to work would have more COVID-19 cases and deaths, (3) as mean earnings increase for a county, average COVID-19 cases and deaths decrease.

This paper dives into our variable selection, data sourcing and cleaning, methods for the regression model, our findings, impact, and limitations of our analysis. We find that health insurance coverage is positively associated with COVID-19 cases and deaths, that counties with a higher percentage of essential workers are positively associated with an increase in cases and deaths. Finally, more individuals commuting and an increase in mean earnings are both associated with a decrease in cases and deaths. All of these predictors take into account the population at the county-level.

*SIS/CAS, American University
†CAS, American University
‡SPA, American University

2 COVID-19 Driven Disparities

This paper and project came about as a result of the increasing amount of scholarly and peer-reviewed literature that sheds light on the ever-present and visible economic, social, and medical disparities that resulted directly from the COVID-19 pandemic. Among this literary discourse, there is overwhelming agreement that strong disparities are present within the American population as a result of individuals who contracted and/or died from COVID-19. Not only are there existing risk factors that make certain marginalized members of populations more susceptible to impact from contraction of COVID-19 but the harms of contraction are further exacerbated (Abedi et al. 2021). With evidence that a large burden of the brute impacts of COVID-19 are felt by the country's essential workers, low-income populations, and by those who found healthcare inaccessible, our research question and variable selection is informed with this insight in mind (Kantamneni 2020). Even in some of the more basic health instructions administered by the United States government during the COVID-19 pandemic, such as social distancing, there are severe disparities by levels of commute and income level in individuals being able to adhere to those restrictions (Garnier et al. 2021). From the American Community Survey, one the country's leading resources for measuring impacts among populations in the United States, we select economic, social, and health indicators to further evaluate the nature of the impact that COVID-19 truly had within the population of the United States.

With a large portion of the scholarly conversation surrounding the interdisciplinary human impact of COVID-19 being centered around the United States, we decide to take a micro approach to see its specific impact in our communities of DC, Maryland, and Virginia.

3 Data and Methods

3.1 Data Sources

This project sources two data sets 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 the US Census Bureau's API and COVID-19 Data is acquired from Covid ActNow's API.

Our final combined data set, with both ACS and COVID-19 data, has 143 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 the American 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 is 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 want to investigate from the American Community Survey, we are 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 independent 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 start out with time series COVID-19 data for Maryland, DC, and Virginia. We only select the year 2021 for COVID-19 data in order to match the ACS data from 2021. We are able to get a daily cumulative case and death count from COVID ActNow, but in order to ensure that our analysis is on the same level as ACS data, we create two new variables – average cases from 2021 and average deaths from 2021. The specific dependent variables of interest from this data set are:

- 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 combine both data sets to prepare it for analysis. We are able to load in both Census ACS data and COVID-19 data from COVID ActNow through their respective APIs, and then subsequently clean the source data.

For cleaning the COVID-19 data, we parse out DC, Maryland, Virginia from our complete time series data from COVID ActNow, then create a separate data set with just these three geographic locations, and compute average cases and deaths for each county during 2021. Then we merge both data sets to combine variables from both individual source sets. Finally, we recalculate all the variables as a rate based on the 2021 population for each county (population data was gathered from the US Census).

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 is aided by the `tidycensus` package, which allows us to pull direct data from the ACS/US Census Bureau. Data analysis/visualization is conducted using `ggplot2` from the Tidyverse, as well as `stargazer` to create tables (Hlavac 2018). Finally, we use the `keyring`, `httr`, and `jsonlite` packages for API imports from the respective sources.

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.

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 death and case rates. This finding seems counter-intuitive and was something we wanted to examine more closely after controlling for the effect of other predictors.

This exploratory data analysis leads us to pursue linear regression as a method for analyzing multiple economic indicators and mean COVID outcomes. Specifically, we arrive 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

We begin by estimating an Ordinary Least Squares (OLS) model that predicts average cases and deaths (in 2021) for all counties in DC, Maryland, and Virginia. In our model, we consider: the percent of workers that could be considered essential, the number of individuals that have health insurance, the number of workers that commute to work, and the mean individual earnings. Since state and local vaccination efforts were underway for the duration of 2021, we also control for average vaccinations administered. The ultimate goal of our analysis is to ascertain which economic indicators were significant in predicting COVID outcomes.

Overall, our models explain about 30% of variation in both average cases and deaths for all counties in DC, MD, VA in 2021. The following paragraphs will dive deeper into each model and elucidate the significance of our results.

Table 1: Two Regression Models Predicting Variation in 2021 COVID Outcomes

	For all counties in DC, MD, VA	
	Average Cases	Average Deaths
Percent Essential Workers	20.795^{***} (4.602)	0.408^{**} (0.172)
Number of Individuals with Health Insurance	0.096^{***} (0.032)	0.005^{***} (0.001)
Number of Individuals Commuting to Work	-0.108^{**} (0.051)	-0.007^{***} (0.002)
Mean Earnings (USD)	-0.005^{***} (0.001)	-0.0001 (0.0001)
Average Vaccinations Administered	-0.007 (0.015)	-0.001 (0.001)
Intercept	0.049^{***} (0.017)	0.0005 (0.001)
Corrected AIC	-537.5	-1359.5
R^2	0.308	0.317
F Statistic (df = 5; 119)	10.578^{***}	11.046^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

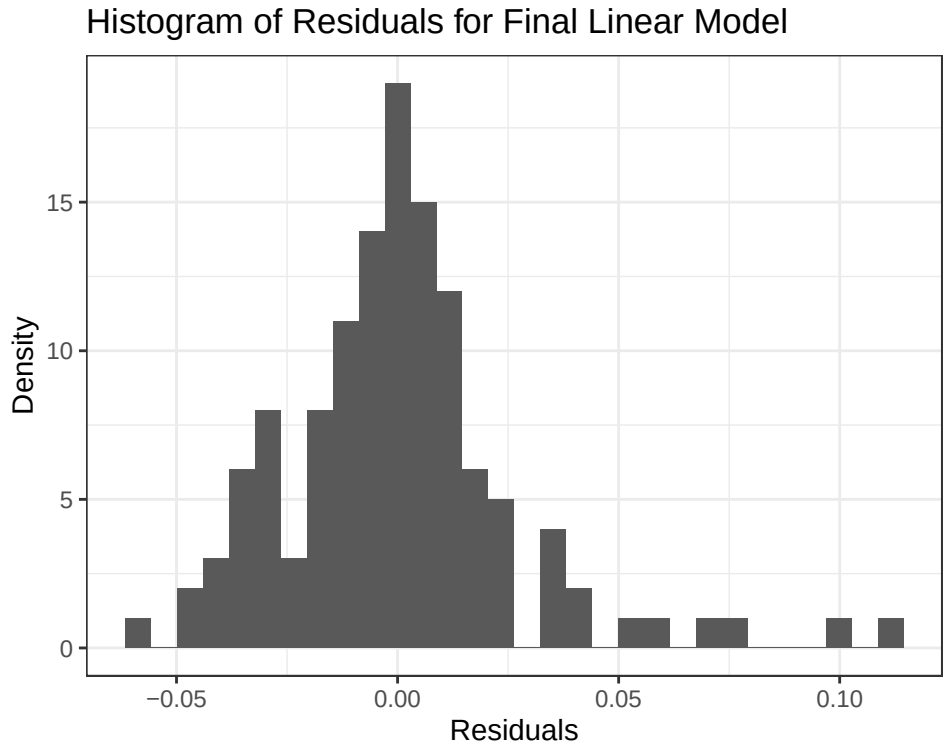
4.1 Average Cases

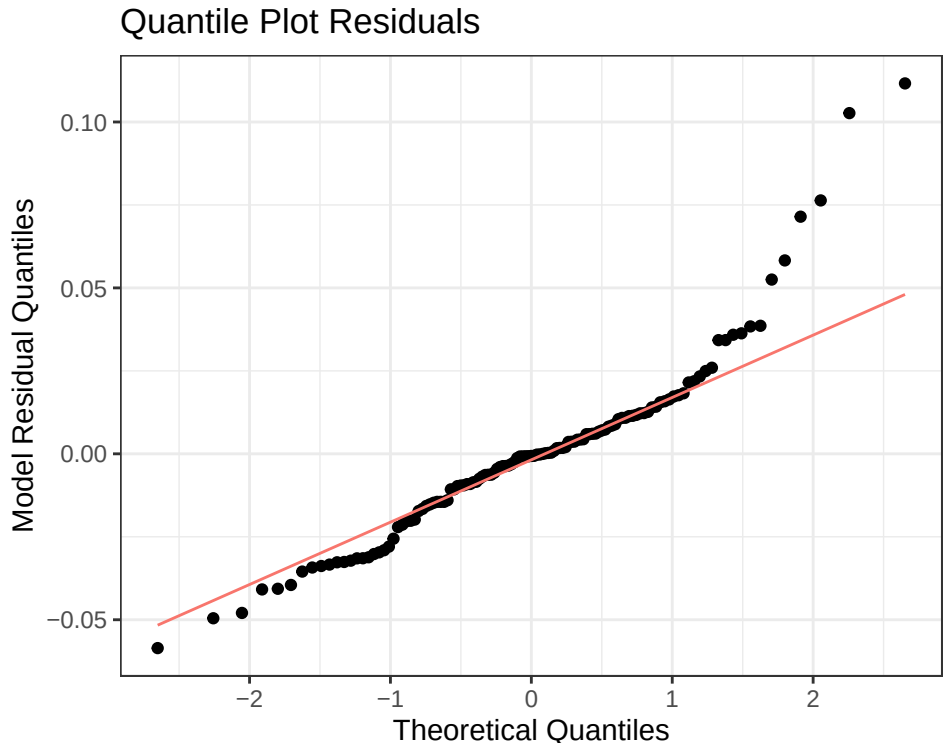
Our baseline model explains about 30% of the variance in average cases. In our analysis, every variable except average vaccinations was significant at least at $\alpha = .05$.

$$\widehat{Average\ Cases}_i = \beta_0 + \beta_1(Percent\ Essential\ Workers_i) + \beta_2(Number\ of\ Individuals\ with\ Health\ Insurance_i) + \beta_3(Average\ Vaccinations\ Administered_i) + \beta_4(Mean\ Earnings_1) + \beta_5(Number\ of\ Individuals\ Commuting\ to\ Work_i) + \epsilon_i$$

Notably, the percent of essential workers in a given county has by far the largest effect on average cases. Indeed, for every additional percent of essential workers in a county, average cases will increase by about 20. The estimates for number of individuals with health insurance and is nearly zero. For every additional worker commuting to work we predict a decrease in cases by about .108. Finally, for a decrease in mean earnings of .005 (USD) we predict on more case. Throughout this analysis we control for average vaccinations administered (throughout the year).

The model assumptions are generally satisfactory. A histogram of the residuals shows that the residuals follow a normal distribution despite some outlying values on both the upper and lower tails. This is also confirmed by the normal Q-Q plot which shows that most of the residuals follow a normal pattern.





Our data has multiple extreme outlying values, however, only two have a relatively high leverage and only seven have a Cooks' Distance above our threshold of three times the mean of all distances. Thus, of our

sample of 143 observations only seven are extreme and/or influential. After dropping these values and re-running our model our test statistic did not change drastically. As a result we retained these values in our final model.

Cooks' Distances Above Threshold
0.15140295
0.05371456
0.09227393
0.16444474
0.05071478
0.04845231
0.08283901
0.69817222

In the next section, we will repeat our analysis against average deaths using the same predictors.

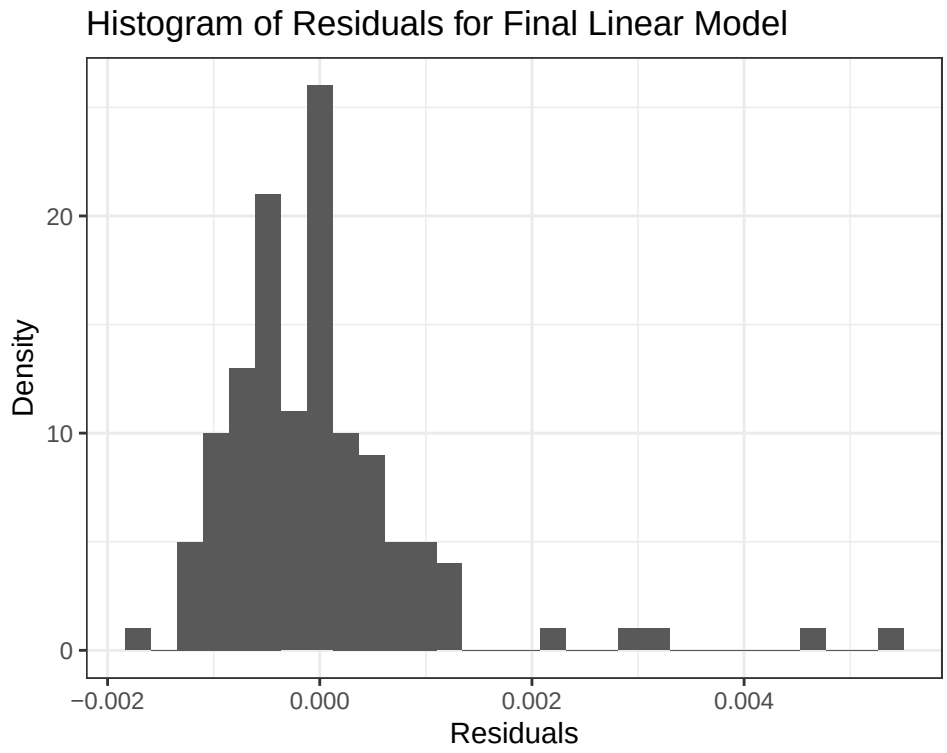
4.2 Average Deaths

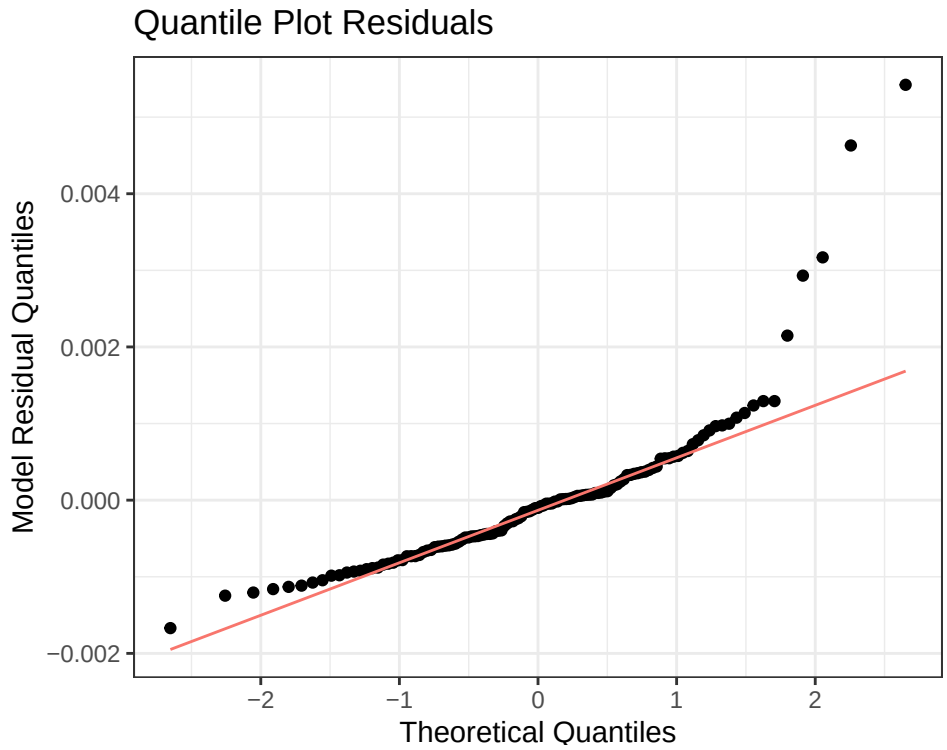
Our baseline model explains about 31% of the variance in average deaths. Three predictors were significant at $\alpha = .01$. Our final model is estimated below:

$$\widehat{Average\ Deaths} = \beta_o + \beta_1(Percent\ Essential\ Workers) + \beta_2(Number\ of\ Individuals\ with\ Health\ Insurance) + \beta_3(Number\ of\ Individuals\ Commuting\ to\ Work) + \beta_4(MeanEarnings) + \beta_5(Average\ Vaccinations\ Administered) + e_i$$

Between the two models, it is unclear if this model performs better because of comparable R-sq and lower AIC. Notably, we find that for every additional percent in essential workers we estimate average deaths to increase by about .408, while holding other variables constant. We also find that as mean earnings increases, we can expect average deaths to decrease. However, the effect of this is extremely negligible. For every additional individual with health insurance, we predict deaths to increase by about .005. We examine this relationship more in the next section. We also find that as the number of individuals commuting to work increases by one, average deaths decrease by .007.

In checking the model assumptions we again see evidence of outlying values, however, for the most part the residuals are normally distributed. In this model, only seven observations had Cooks' Distances above our threshold. Again, we retained these values due to insufficient changes to the test statistic after dropping them.





Cooks' Distances Above Threshold
0.03949423
0.05338901
0.06391039

	0.08382036
	0.38163565
	0.29976403
	0.24550182

5 Discussion

The COVID-19 pandemic has renewed conversations about health equity disparities. As data becomes more abundant, literature has emerged seeking to understand health outcomes, policy interventions, and disparities (Thornton and Yang 2023). This research project aims to contribute to this growing body of literature by examining the relationship between socio-economic factors and mean COVID outcomes for all counties/wards in Virginia, Maryland, and the District of Columbia.

Notably, we find evidence that suggests that counties with a higher percent of essential workers are predicted to have more cases and deaths on average. The effect, however, is substantially greater for average cases. This finding is consistent with what we would expect, as well as existing theoretical explanations. Indeed, Rosengren et al. (2022), Rhodes et al. (2022), Do and Frank (2021), and Welsh (2022) all underscore the elevated risks that essential workers face(d) during the pandemic and outbreaks. Our analysis, however, considers the larger implications of this by estimating how many more cases and deaths a local county could possibly expect based on the percentage of essential workers. While these findings cannot be generalized beyond the mid-Atlantic – and specifically DC, Maryland, and Virginia – it does provide and important contribution in measuring this phenomenon at the local level.

We also find a puzzling positive association between worse COVID deaths and health insurance coverage. This finding – while strange – is not novel Cuadros et al. (2023) document a similar association between insurance coverage and a higher burden of disease. Importantly, the authors note that there isn't a "negative impact of insurance on health status" – rather a series of factors which may be at play. First, insured individuals may be more likely to seek care than uninsured individuals – something which is well documented in the consistent and pervasive under-diagnosis of diseases/conditions in the uninsured population. Second, population size may play an intervening role. Larger counties/locales which are predisposed to having more cases could plausibly have a greater raw number of individuals who are insured. In short, we don't place much emphasis on the effects of health insurance. The literature over overwhelmingly suggests that health insurance plays a positive role in health status Uninsurance (2002), however, statistical results like ours are possible Cuadros et al. (2023) for a variety of reasons. Thus, these results likely indicate there are other factors at play that our model doesn't adequately capture.

Our final research question is aimed at understanding the role of earnings on average COVID outcomes. We found that as mean earnings increase, counties could expect less cases and deaths on average. This measure should not be generalized beyond DC, Maryland, and Virginia and should be used with caution. The relationship between wealth and health disparities has been well-documented in both the COVID-19 pandemic and beyond (Tai et al. 2022; Khayat, Teron, and Rasoulyan 2022). However, there are likely many variables which we did not include due to project time-constraints, that would enhance our estimate. Specifically, mean earnings vary by location, industry, and population size – all things which future research should consider as factors when estimating the relationship between earnings and health outcomes. Despite this, our estimate contributes to this growing research body and provides a baseline for future analysis.

Our research aims at exploring the intersection of economic indicators and average COVID outcomes. To date, much research has been done on the role of socio-economic factors and health outcomes. Our research builds on this by examining this for three states in the mid-Atlantic. Overall, we discover that the percent of essential workers and mean earnings (at the county level) were important variables when estimating average COVID outcomes. Our models explain about 30% of the variation in average COVID outcomes within our sample. Moving forward, future research should be aimed at expanding both the temporal and spatial scopes we employed as well as consider more socio-economic variables which we didn't have the chance to incorporate. Additionally, future research should consider other methods – such as an ordinal probit model

– which may do a better job at modeling the relationship between socio-economic indicators and health outcomes.

References

Abedi, Vida, Oluwaseyi Olulana, Venkatesh Avula, Durgesh Chaudhary, Ayesha Khan, Shima Shahjouei, Jiang Li, and Ramin Zand. 2021. "Racial, Economic, and Health Inequality and COVID-19 Infection in the United States." *Journal of Racial and Ethnic Health Disparities* 8 (3): 732–42. https://doi.org/10.1007/s40615-020-00833-4.

Cuadros, D. F., J. D. Gutierrez, C. M. Moreno, S. Escobar, F. D. Miller, G. Musuka, R. Omori, P. Coule, and N. J. MacKinnon. 2023. "Impact of Healthcare Capacity Disparities on the COVID-19 Vaccination Coverage in the United States: A Cross-Sectional Study." *Lancet Regional Health - Americas* 18. https://doi.org/10.1016/j.lana.2022.100409.

Do, D. P., and R. Frank. 2021. "U.S. Frontline Workers and COVID-19 Inequities." *Preventive Medicine* 153. https://doi.org/10.1016/j.ypmed.2021.106833.

Garnier, Romain, Jan R Benetka, John Kraemer, and Shweta Bansal. 2021. "Socioeconomic Disparities in Social Distancing During the COVID-19 Pandemic in the United States: Observational Study." *Journal of Medical Internet Research* 23 (1): e24591. https://doi.org/10.2196/24591.

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.

Kantamneni, Neeta. 2020. "The Impact of the COVID-19 Pandemic on Marginalized Populations in the United States: A Research Agenda." *Journal of Vocational Behavior* 119 (June): 103439. https://doi.org/10.1016/j.jvb.2020.103439.

Khayat, F., L. Teron, and F. Rasoulyan. 2022. "COVID-19 and Health Inequality: The Nexus of Race, Income and Mortality in New York City." *International Journal of Human Rights in Healthcare* 15 (4): 363–72. https://doi.org/10.1108/IJHRH-05-2021-0110.

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/.

Rhodes, S., J. Wilkinson, N. Pearce, W. Mueller, M. Cherrie, K. Stocking, M. Gittins, S. V. Katikireddi, and M. V. Tongeren. 2022. "Occupational Differences in SARS-CoV-2 Infection: Analysis of the UK ONS COVID-19 Infection Survey." *Journal of Epidemiology and Community Health* 76 (10): 841–46. https://doi.org/10.1136/jech-2022-219101.

Rosengren, A., M. Söderberg, C. E. Lundberg, M. Lindgren, A. Santosa, J. Edqvist, M. Åberg, et al. 2022. "COVID-19 in People Aged 18–64 in Sweden in the First Year of the Pandemic: Key Factors for Severe Disease and Death." *Global Epidemiology* 4. https://doi.org/10.1016/j.gloepi.2022.100095.

Tai, D. B. G., I. G. Sia, C. A. Doubeni, and M. L. Wieland. 2022. "Disproportionate Impact of COVID-19 on Racial and Ethnic Minority Groups in the United States: A 2021 Update." *Journal of Racial and Ethnic Health Disparities* 9 (6): 2334–39. https://doi.org/10.1007/s40615-021-01170-w.

Thornton, R. L. J., and T. J. Yang. 2023. "Addressing Population Health Inequities: Investing in the Social Determinants of Health for Children and Families to Advance Child Health Equity." *Current Opinion in Pediatrics* 35 (1): 8–13. https://doi.org/10.1097/MOP.0000000000001189.

Uninsurance, Institute of Medicine (US) Committee on the Consequences of. 2002. *Effects of Health Insurance on Health.* National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK220636/.

Welsh, C. E. 2022. "Workplace Contact Patterns Across Occupational Groups." *The Lancet Regional Health - Europe* 16. https://doi.org/10.1016/j.lanepe.2022.100356.

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.